DOCUMENT RESUME

ED 395 964 TM 025 102

AUTHOR Holland, Paul W.; Thayer, Dorothy T.
TITLE The Kernel Method of Equating Score Distributions. Program Statistics Research Technical Report No. 89-84.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-89-7
PUB DATE Feb 89
NOTE 58p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Data Collection; *Equated Scores; *Estimation (Mathematics); *Nonparametric Statistics; Research Design; *Statistical Distributions
IDENTIFIERS Anchor Tests; Equipercentile Equating; *Kernel Method; Log Linear Models; Smoothing Methods

ABSTRACT

A new and unified approach to test equating is described that is based on log-linear models for smoothing score distributions and on the kernel method of nonparametric density estimation. The new method contains both linear and standard equipercentile methods as special cases and can handle several important equating data collection designs. An example is used to illustrate the new method for the random groups and external anchor-test designs. The kernel method of equating, when coupled with estimated score distributions using log-linear models, has a number of advantages over other observed-score equating methods. The three phases of estimation, continuization, and equating form a unified approach to many equating problems. The kernel method contains linear and traditional equipercentile methods as special cases and can exploit the best features of both methods. An appendix presents a proof of the paper's second theorem. (Contains 7 tables, 9 figures, and 12 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 395 964

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

RR-89-7

The Kernel Method of Equating Score Distributions

Paul W. Holland
Dorothy T. Thayer

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT NO. 89-84

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541

BEST COPY AVAILABLE

TM025102

THE KERNEL METHOD OF EQUATING SCORE DISTRIBUTIONS

Paul W. Holland

Dorothy T. Thayer

Program Statistics Research
Technical Report No. 89-84

Research Report No. 89-7

Educational Testing Service
Princeton, New Jersey 08541-0001

February 1989

Copyright © 1989 by Educational Testing Service. All rights reserved.

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

TABLE OF CONTENTS

		Page
1.	Introduction	1
2.	The Kernel Method of Equating	7
3.	The Estimation Step	15
	3.1 Random Groups Equating Design	17
	3.2 The Anchor-Test Equating Design	21
4.	The Continuization Step	30
5.	The Equating Step	40
6.	Discussion	43
	References	46
	Appendix	47

ABSTRACT

A new and unified approach to test equating is described that is based on log-linear models for smoothing score distributions and on the kernel method of non-parametric density estimation. The new method contains both linear and standard equipercentile methods as special cases and can handle several important equating data collection designs. An example is used to illustrate the new method for the random groups and external anchor-test designs.

1. INTRODUCTION

This paper introduces a new and unified approach to test equating based on a flexible family of equating functions that contains both the linear and the equipercentile equating functions as special cases. The new method grows out of the perspective on observed-score test equating described in Braun and Holland (1982). We call the new approach the "kernel method of equating tests" because of its close connection to the well-studied methods of non-parametric density estimation using a gaussian kernel, Tapia and Thompson (1978). The kernel method may be viewed as generalizing certain features of the equipercentile method described by Angoff (1984). Because of this we first review the equipercentile method from our perspective; this also allows us to introduce our notational scheme.

Review of equipercentile equating

Suppose we have two tests, denoted by X and Y, and let the possible raw-score values for X and Y be denoted by $x_1,\ldots,x_J$ and $y_1,\ldots,y_K$, respectively. In this notation, J and K are the number of possible raw-score values and not the number of test items on X and Y. In the applications that concern us, $x_1,\ldots,x_J$ will denote consecutive integers; similarly for $y_1,\ldots,y_K$. If, for example, X is a number-right scored test, then $x_1 = 0$, $x_2 = 1,\ldots$ and x_J = the number of items in test X. Alternatively, for a rounded formula-scored X, x_1 is negative but x_J still denotes the number of items in X.

As Braun and Holland (1982) emphasize, observed-score test equating always takes place on a specific population of examinees. We suppose that this population is fixed and let r_j and s_k denote the score probabilities for this population, i.e.,

BEST COPY AVAILABLE

$$r_j = \text{Prob}\{X = x_j\} \quad , \quad s_k = \text{Prob}\{Y = y_k\} \quad . \tag{1}$$

In (1), we abuse notation slightly and let X denote both a test and the score of a randomly selected examinee on this test. (Similarly for Y). The score probabilities, $\{r_j\}$ and $\{s_k\}$, are population parameters and depend on the underlying population of examinees. They must be estimated from the data collected in the equating experiment. We defer a serious discussion of how they might be estimated to section 3 and merely suppose that estimates, $\{\hat{r}_j\}$ and $\{\hat{s}_k\}$, are available.

Associated with the score probabilities are the cumulative distribution functions (cdfs) of the test scores for X and Y that are defined by

$$F(x) = \text{Prob}(X \leq x) = \sum_{\substack{j \\ x_j \leq x}} r_j \ , \tag{2}$$

and

$$G(y) = \text{Prob}(Y \leq y) = \sum_{\substack{k \\ y_k \leq y}} s_k \ . \tag{3}$$

In (2), x denotes any real number and the summation is over all j for which x_j does not exceed x. In (3), y denotes any real number and the sum is over all k for which y_k does not exceed y. The cdfs, F and G, defined in (2) and (3) are step functions with jumps at the possible values for X and Y, respectively.

If F and G were continuous cdfs (as is, for example, the cdf for the normal distribution) then the equipercentile equating function for equating X to Y would have the form

$$e_Y(x) = G^{-1}(F(x)) \tag{4}$$

and for equating Y to X it would have the form

$$e_X(y) = F^{-1}(G(y)) \tag{5}$$

where F^{-1} and G^{-1} denote the inverse functions of F and G defined by

$$x = F^{-1}(p) \text{ if and only if } p = F(x)$$

and

$$y = G^{-1}(p) \text{ if and only if } p = G(y).$$

See Braun and Holland (1982) for more discussions of this description of equipercentile equating.

If F and G were continuous, the function $e_Y(x)$ and $e_X(y)$ defined in (4) and (5) would exactly match the distribution of $e_X(Y)$ to that of X and the distribution of $e_Y(X)$ to that of Y. However, in practice F and G are discrete so that, strictly speaking, F^{-1} and G^{-1} do not exist and hence e_Y and e_X cannot be defined as in (4) and (5). This fact is usually glossed over in discussions of equipercentile equating (e.g. Angoff, 1984; Lord, 1950). Instead, F and G are approximated by linear interpolation to obtain percentile ranks. It is instructive to see exactly how this linear interpolation is derived mathematically, and we now do this.

The <u>percentile rank</u> of a score x_k is defined as the proportion of examinees in the population scoring below x_k plus one-half of the proportion scoring exactly x_k (Angoff, 1984). How can such a definition be justified? Here is one approach to justifying it.

Suppose U is a random variable with a uniform distribution on $(-\frac{1}{2},\frac{1}{2})$, and suppose that U is independent of the discrete random variable X where

$$r_j = \text{Prob}\{X = x_j\},\ j = 1,\ldots,J.$$

The cdf of U is given by

$$\text{Prob}\{U \le u\} = \begin{cases} 1 & \text{if } u \ge \frac{1}{2}, \\ 0 & \text{if } u \le -\frac{1}{2}, \\ u + \frac{1}{2} & \text{if } -\frac{1}{2} \le u \le \frac{1}{2}. \end{cases} \qquad (6)$$

Now consider a new random variable $X_\star$ defined by

$$X_\star = X + U \ . \tag{7}$$

The new variable $X_\star$ has a continuous distribution that is spread over the interval $x_1 - \frac{1}{2}$ to $x_J + \frac{1}{2}$. The cdf of $X_\star$ is found as follows.

$$F_\star(x) = \text{Prob}\{X_\star \leq x\} = \text{Prob}\{X + U \leq x\}$$

$$= \sum_j \text{Prob}\{X + U \leq x \mid X = x_j\}\text{Prob}\{X = x_j\} = \sum_j \text{Prob}\{U \leq x - x_j \mid X = x_j\}\, r_j$$

$$= \sum_j \text{Prob}\{U \leq x - x_j\}\, r_j \ .$$

But from (6) it follows that

$$\text{Prob}\{U \leq x - x_j\} = \begin{cases} 1 & \text{if } x \geq x_j + \frac{1}{2} \ , \\ 0 & \text{if } x \leq x_j - \frac{1}{2}, \\ x - x_j + \frac{1}{2} & \text{if } x_j - \frac{1}{2} \leq x \leq x_j + \frac{1}{2} \ , \end{cases} \tag{8}$$

and hence we have

$$F_\star(x) = \sum_{x_j \leq x - \frac{1}{2}} r_j + (x - x_i + \tfrac{1}{2})r_i, \quad \text{for } x_i - \tfrac{1}{2} \leq x \leq x_i + \tfrac{1}{2}, \tag{9}$$

where the summation in (9) is over all j for which x_j does not exceed $x - \frac{1}{2}$. Now evaluate $F_\star(x)$ at x_i and we have

$$F_\star(x_i) = \sum_{x_j < x_i} r_j + \tfrac{1}{2} r_i \ , \tag{10}$$

which is the probability of scoring below x_i plus one half the probability of scoring exactly x_i and this is the definition of percentile ranks given above. This shows that the percentile rank of x_i is simply the value of the cdf $F_\star$ at

x_i, i.e. $F_\star(x_i)$. We may view $F_\star$ as a continuous approximation to the step-function F. From (9) we see that $F_\star$ is a piecewise linear function that starts at zero at $x_1 - \frac{1}{2}$ and (if the x_j are consecutive intergers and $r_j > 0$) steadily increases to the value of 1 at $x_J + \frac{1}{2}$.

The standard version of equipercentile equating can be viewed as replacing F by $F_\star$ and G by a corresponding $G_\star$. When $\{r_j>0\}$ and $\{s_k>0\}$, the inverse functions, $F_\star^{-1}$ and $G_\star^{-1}$ both exist and the functions in (11) are well-defined, i.e.,

$$e_Y(x) = G_\star^{-1}(F_\star(x))$$

and (11)

$$e_X(y) = F_\star^{-1}(G_\star(y)) \ .$$

By definition, e_Y and e_X given in (11) are the population equipercentile equating functions for equating X and Y. Sample estimates of e_Y and e_X in (11) are defined by substituting in $\hat{r}_j$ for r_j and $\hat{s}_k$ for s_k in the definitions of $F_\star$ and $G_\star$, i.e. (9). (In addition, in practice a post-smoothing step may be introduced to make the final equating functions even smoother than the piecewise linear functions in (11), Angoff (1984), Fairbank (1985), Kolen (1984), Kolen and Jarjoura (1987)).

There are various problems with this version of equipercentile equating. For one, consider the mean and variance of X and its "continuous approximation", $X_\star$. We have

$$E(X_\star) = E(X + U) = E(X) + 0 = E(X)$$

but,

$$Var(X_\star) = Var(X + U) = Var(X) + Var(U) \ .$$

It is well-known that $Var(U) = 1/12$ so that X and $X_\star$ have the same means but different variances. The higher moments of $X_\star$ also fail to agree with those of

X. Hence, what the traditional version of equipercentile equating actually does is to exactly match the distribution of the two continuous random variables $X_\star$ and $Y_\star$ rather than to match the discrete distributions of X and Y. No moments beyond the first can be expected to be exactly matched using the standard equipercentile equating function although they may be close enough for practical work. In addition, because $F_\star$ and $G_\star$ place no probability outside the intervals $(x_1 - \frac{1}{2}, x_J + \frac{1}{2})$ and $(y_1 - \frac{1}{2}, y_K + \frac{1}{2})$, it is automatically true that $e_Y(x)$ and $e_X(y)$ defined in (11) map the end-points of these two intervals onto each other. This is often an undesirable property in test equating since it usually forces the highest (and lowest) score on X to be mapped onto the highest (and lowest) score on Y. If X were much easier than Y this property is unreasonable and is due solely to the arbitrary use of $F_\star$ and $G_\star$ to form the equating functions.

These problems with the traditional form of equipercentile equating all stem from the arbitrary form assumed for U, i.e. that it be uniform on $(-\frac{1}{2}, \frac{1}{2})$. The crux of the kernel method is to replace U with a more flexible choice of random variable. In particular, the point of view taken here is that the traditional equipercentile method is a version of the kernel method using a fixed "bandwidth" (i.e. the variance of continuous random variable added to X). In general, it is always better to use bandwidths that can vary in useful ways when kernel methods are employed.

2. THE KERNEL METHOD OF EQUATING

Our approach is to accept the fact that X and Y are discrete and hence that F and G must be approximated, in some sense, by continuous cdfs before (4) and (5) can become well-defined (as they are in (11)). Picking up on the ideas in section 1, suppose we now consider the distribution of the random variable, $X(h_X)$, defined by

$$X(h_X) = a_X(X + h_X V) + (1-a_X)\,\mu_X \tag{12}$$

where X is the discrete random variable that appeared in section 1 and V is a random variable that is independent of X and has a standard normal, N(0,1), distribution. Also, in (12) μ_X and a_X are defined by:

$$\mu_X = E(X) = \sum_j x_j r_j \ , \tag{13}$$

$$a_X^2 = \frac{\sigma_X^2}{\sigma_X^2 + h_X^2} \ , \tag{14}$$

$$\sigma_X^2 = \mathrm{Var}(X) = \sum_j (x_j - \mu_X)^2 r_j \ . \tag{15}$$

The bandwidth, h_X, is a non-negative constant that we are free to select to achieve some useful purpose. What we have done in (12) is replaced U in (7) by $h_X V$ and then rescaled the sum of X and $h_X V$ to preserve the mean and variance of X, i.e. it is easy to show that

$$E(X(h_X)) = E(X) = \mu_X$$

and

$$\mathrm{Var}(X(h_X)) = \mathrm{Var}(X) = \sigma_X^2 \ .$$

for any choice of $h_X \geq 0$. Observe that X(0) is identical to X and X(∞) is a normal random variable with the same mean and variance as X. When $h_X > 0$, $X(h_X)$ has a continuous distribution with cdf

$$F_{h_X}(x) = \text{Prob}\{X(h_X) \leq x\} \quad . \tag{16}$$

We will regard $\{F_{h_X}(x)$, for $h_X > 0\}$, as a family of continuous approximations to the discrete cdf F(x). Hence, instead of the single X_* of section 1, we may consider the entire collection of approximations, $\{X(h_X)$, $h_X > 0\}$.

Observe that $\text{Var}(h_X V) = h_X^2$, whereas in section 1, Var(U) = 1/12. Hence

$$h_X = 1/\sqrt{12} = .289 \sim .3 \tag{17}$$

corresponds roughly to the traditional form of the continuous approximation to F used in equipercentile equating, i.e. $F_*(x)$ in (9).

A nice feature of $F_{h_X}(x)$ is that it has a reasonably tractable analytic form. This is given in theorem 1, below.

Theorem 1: If $X(h_X)$ is defined by (12) and $F_{h_X}(x)$ is the cdf in (16) then

$$F_{h_X}(x) = \sum_j r_j \, \Phi(R_{jX}(x)) \tag{18}$$

where $\Phi(x)$ denote the standard normal cdf and $R_{jX}(x)$ is the linear function of x given by

$$R_{jX}(x) = \frac{x - a_X x_j - (1-a_X)\mu_X}{a_X \; h_X} \quad . \tag{19}$$

In (19), a_X and μ_X are defined as in (13) - (15).

Proof:

$$F_{h_X}(x) = \text{Prob}\{X(h_X) \le x\} = \text{Prob}\{a_X(X + h_X V) + (1-a_X)\mu_X \le x\}$$

$$= \text{Prob}\{a_X h_X V \le x - a_X X - (1-a_X)\mu_X\}$$

$$= \sum_j \text{Prob}\{a_X h_X V \le x - a_X x_j - (1-a_X)\mu_X | X = x_j\} r_j$$

$$= \sum_j \text{Prob}\left\{V \le \frac{x - a_X x_j - (1-a_X)\mu_X}{a_X h_X}\right\} r_j = \sum_j r_j \Phi(R_{jX}(x)) . \qquad \text{QED.}$$

Because the mean and variance of $X(h_X)$ exactly match those of the original discrete random variable X, it is of interest to know how the higher moments of $X(h_X)$ differ from those of X. It is, however, the cumulants of $X(h_X)$ rather than its moments that have the simplest relationship to those of X. The j^{th} cumulant of a distribution is the coefficient of $(t)^j/j!$ in the Taylor expansion (about zero) of the natural logrithm of its moment generating function, M(t). It is well-known that the first and second cumulants are the mean and variance, respectively, of the distribution. Furthermore, the third and higher cumulants of any normal distribution are all zero. See Kendall and Stuart (1958) for a thorough discussion of cumulants.

Theorem 2 shows the relationship between the cumulants of $X(h_X)$ and those of X.

Theorem 2: If $k_j(h_X)$ denotes the j^{th} cumulant of $X(h_X)$, and k_{jX} denotes the j^{th} cumulant of X, then for $j \ge 3$ we have

$$k_j(h_X) = (a_X)^j k_{jX} , \qquad (20)$$

where a_X is defined in (14).

The proof of Theorem 2 is given in the appendix.

BEST COPY AVAILABLE

We may interpret Theorem 2 by saying that the higher cumulants of $X(h_X)$ are all smaller in absolute size (i.e. more like those of the normal distribution) than the corresponding cumulants of the original distribution of X. This is because

$$(a_X)^j < 1 \text{ if } h_X > 0. \tag{21}$$

The kernel method of equating is now easy to describe. First of all, continuous approximations to F and G are found via (18), i.e.,

$$F_{h_X}(x), \text{ and } G_{h_Y}(y), \tag{22}$$

and then the equating functions $e_Y(x)$ and $e_X(y)$ are defined by

$$e_Y(x) = G^{-1}_{h_Y}(F_{h_X}(x)) \tag{23}$$

and

$$e_X(y) = F^{-1}_{h_X}(G_{h_Y}(y)). \tag{24}$$

Note that (23) and (24) define <u>families</u> of equating functions indexed by h_X and h_Y.

In (23) and (24) the inverse functions $F^{-1}_{h_X}$ and $G^{-1}_{h_Y}$ are defined by

$$x = F^{-1}_{h_X}(p) \text{ if and only if } p = F_{h_X}(x) \text{ and}$$

$$y = G^{-1}_{h_Y}(p) \text{ if and only if } p = G_{h_Y}(y).$$

In practice, these inverse functions do not have an explicit form but they can be easily computed by interpolation.

In (22) the bandwidths, h_X and h_Y, are called the continuization constants. When they are both chosen to be .3 the resulting equating functions, (23) and (24), agree closely with the traditional equipercentile equating functions, as noted in (17). When h_X and h_Y are both large, the equating functions closely appoximate the standard linear equating function as we demonstrate in the next theorem.

Theorem 3: If $e_Y(x)$ is defined by (23) then

$$\lim_{h_X, h_Y \to \infty} e_Y(x) = \mu_Y + \frac{\sigma_Y}{\sigma_X}(x - \mu_X) = \mathrm{Lin}_Y(x) .$$

Proof: It is obvious that as h_X and $h_Y \to \infty$, $F_{h_X}(x)$ and $G_{h_Y}(y)$ approach these normal cdf's:

$$F_{h_X}(x) \to \Phi\left(\frac{x - \mu_X}{\sigma_X}\right) ,$$

and

$$G_{h_Y}(y) \to \Phi\left(\frac{y - \mu_Y}{\sigma_Y}\right) .$$

Hence

$$G^{-1}_{h_Y}(p) \to \mu_Y + \sigma_Y \Phi^{-1}(p) ,$$

where $\Phi^{-1}(p)$ is the inverse of the standard normal cdf, therefore

$$e_Y(x) \to \mu_Y + \sigma_Y \Phi^{-1}\left(\Phi\left(\frac{x - \mu_X}{\sigma_X}\right)\right)$$

$$= \mu_Y + \sigma_Y \left(\frac{x - \mu_X}{\sigma_X}\right) . \qquad \text{QED}$$

We now point out that the objections to equipercentile equating mentioned at the end of section 1 do not apply to the kernel method of equating. By varying the choice of the continuization constants, h_X and h_Y, we may achieve a wide variety of equating functions that are "in between" the traditional linear and equipercentile functions. All of these equating functions exactly match the means and variances of $e_X(Y)$ and X and of $e_Y(X)$ and Y. Furthermore, depending on the choice of continuization constants the equating functions need not map the top and bottom scores on X onto the top and bottom scores of Y. Moreover, the equating functions given by (23) and (24) are defined for all x and y and are not restricted to the raw score intervals, i.e. $[x_1, x_J]$ and $[y_1, y_K]$.

Summary of the kernel method of equating

We view observed-score test equating as having three distinct steps or phases, each of which involve separate ideas.

Phase 1: The estimation step. In this step, estimates of $\{r_j\}$ and $\{s_k\}$ are obtained, $\{\hat{r}_j\}$ and $\{\hat{s}_k\}$. This is a purely statistical phase in which various models for the data are tried out and are selected to give a good fit to the data. These models then generate the values of $\{\hat{r}_j\}$ and $\{\hat{s}_k\}$. We recommend that log-linear models like those described in Holland and Thayer (1987) and Rosenbaum and Thayer (1987) be used to do this data fitting since they are flexible enough to describe a wide variety of real situations. In section 3 we illustrate this approach.

Phase II: The continuization step. In this step, h_X and h_Y are chosen to determine continuous approximations, $\hat{F}_{h_X}(x)$ and $\hat{G}_{h_Y}(y)$, to $\hat{F}(x)$ and $\hat{G}(y)$. ($\hat{F}(x)$ and $\hat{G}(y)$ are obtained by substituting $\hat{r}_j$ for r_j in (2) and $\hat{s}_k$ for s_k in (3)). The approximating cdf's have the form

$$\hat{F}_{h_X}(x) = \sum_j \hat{r}_j \, \Phi\left(\frac{x - \hat{a}_X x_j - (1-\hat{a}_X)\hat{\mu}_X}{h_X \hat{a}_X}\right) \tag{25}$$

and

$$\hat{G}_{h_Y}(y) = \sum_k \hat{s}_k \, \Phi\left(\frac{y - \hat{a}_Y y_k - (1-\hat{a}_Y)\hat{\mu}_Y}{h_Y \hat{a}_Y}\right) . \tag{26}$$

In (25) and (26), the estimated quantities, $\hat{a}_X$, $\hat{a}_Y$, $\hat{\mu}_X$, $\hat{\mu}_Y$, are all found by substituting $\hat{r}_j$ for r_j and $\hat{s}_k$ for s_k in (13) - (15). It should be emphasized that continuization is not a statistical procedure so that "optimal" choices of h_X and h_Y cannot be based on optimizing statistical properties such as the estimation of the $\{r_j\}$ or $\{s_k\}$. Rather, in continuization we are attempting to decide which continuous cdf, $\hat{F}_{h_X}(x)$, is "closest" in some appropriate sense to $\hat{F}(x)$. The naive choice of $h_X = 0$ makes $F_{h_X}(x) = F(x)$, but we are then no longer dealing with continuous cdf's and the whole purpose of continuization (i.e. to get unique inverse functions) is defeated. In section 4 we discuss some methods for choosing h_X and h_Y.

Phase III: The equating step. In this step, the estimated equating functions are computed via the formulas

$$\hat{e}_Y(x) = \hat{G}^{-1}_{h_Y}(\hat{F}_{h_X}(x)) \tag{27}$$

and

$$\hat{e}_X(y) = \hat{F}^{-1}_{h_X}(\hat{G}_{h_Y}(y)) . \tag{28}$$

BEST COPY AVAILABLE

Once phases I and II are completed, phase III is straight-forward. However because it is in this phase that the data on tests X and Y are finally combined we identify it as a separate phase. In phase III, we also include the computation of the standard error of equating (the SEE) that measures the accuracy associated with $\hat{e}_X$ and $\hat{e}_Y$. In a companion paper to this one (Holland, King and Thayer, 1988) we give the details of a computation of the SEE that is based on the estimated standard errors for $\hat{r}_j$ and $\hat{s}_k$ that are available if these estimated score probabilities are obtained in a particular way using the log-linear models described in Holland and Thayer (1987).

3. THE ESTIMATION STEP

The population score probabilities, $\{r_j\}$ and $\{s_k\}$ defined in (1), must be estimated from the data collected in the type of equating experiment that is available to the analyst. Angoff (1984) describes a variety of these experiments. In this section we shall be concerned with two major classes of such experiments -- the random groups designs (Angoff's Designs I and II) and the common item or anchor-test designs (Angoff's Designs III and IV). Each class of design is considered in a separate subsection.

The estimation of the score probabilities is a purely statistical problem in the sense that the $\{r_j\}$ and the $\{s_k\}$ are well-defined parameters and hence estimates of these quantities, say $\{\hat{r}_j\}$ and $\{\hat{s}_k\}$, should have desirable statistical properties. Some authors, e.g. Fairbank (1985), refer to the estimation step as "pre-smoothing". While it is true that the estimates, $\{\hat{r}_j\}$ and $\{\hat{s}_k\}$, ought to exhibit appropriate degrees of smoothness, this can be achieved in various ways. There are at least four statistical properties that might be considered in the choice of the estimated score probabilities. These are listed below.

Consistency: As sample sizes increase, the estimates $\hat{r}_j$ and $\hat{s}_k$ ought to converge, in an appropriate sense, to the population values, r_j and s_k.

Positivity: For each possible score value, x_j and y_k, the estimated score probabilities, $\hat{r}_j$ and $\hat{s}_k$, ought to be positive. For most tests, estimating a score probability to be zero is unreasonable.

Stability: Given the sample sizes involved, the deviations of $\hat{r}_j$ from r_j and $\hat{s}_k$ from s_k ought to be as small as possible. Of course these deviations always involve a random element, and the problem is to keep it to a minimum in an appropriate average sense.

Integrity: When possible (as, for example, in the random groups design) the integrity of the sample mean, variance, and possibly other sample moments ought to be preserved in the estimated score distributions, $\{\hat{r}_j\}$ and $\{\hat{s}_k\}$. This means, for example, that $\Sigma\, x_j \hat{r}_j$ and the sample mean for X are equal and that the sample second moment for X and $\Sigma_j\, x_j^2 \hat{r}_j$ are equal as well.

The approach to the estimation step that we favor is to fit a sequence of parametric models to the data and to make appropriate diagnosis of these fitted models until one is found that describes the data well with as few parameters as possible. The log-linear models described in Rosenbaum and Thayer (1987) and in Holland and Thayer (1987) are especially useful in this regard. These models are all well-behaved because they are exponential families of discrete distribution and may be estimated by maximum likelihood using standard iterative techniques. Because these models are exponential families, maximum likelihood estimation forces the equality of some sample and estimated moments. Our experience is that with 3 - 6 parameters these models can adequately describe a wide variety of univariate score distributions. Bivariate distributions, useful for anchor-test equating designs, are also easily estimated using the class of log-linear models. Finally, these models automatically satisfy the positivity and integrity conditions listed above. Careful data analysis using these models also leads to the consistency and stability conditions being satisfied as well.

3.1 Random Groups Equating Designs

In Angoff's Design I class of equating experiments, two independent random samples are drawn from a common population, P, and test X is administered to one sample while test Y is administered to the other. Angoff's Design II is similar except that after each sample has been tested with either X or Y, they also take the other test as well -- i.e. the two groups take both tests but in counter-balanced order. We will ignore the data pooling problem that arises in Design II and merely mention the close connection of this case to Design I to which we now devote our attention.

The raw data that results from the two random samples in Design I may be summarized as two sets of frequencies, i.e. the X-frequencies,

$$n_j = \text{number of examinees with } X = x_j,$$

and the Y-frequencies,

$$m_k = \text{number of examinees with } Y = y_k.$$

The two sample sizes are given by

$$n = \sum_j n_j \text{ and } m = \sum_k m_k .$$

The raw sample proportions $\{n_j/n\}$ and $\{m_k/m\}$ are estimates of the population parameters $\{r_j\}$ and $\{s_k\}$ respectively. However, rarely will the raw sample proportions satisfy the positivity or stability conditions mentioned earlier. Of course, they always satisfy the consistency and integrity conditions, and, when m and n are very large, the raw sample proportions may be acceptable estimates of the population parameters.

Table 1 gives the raw sample frequencies of number-right scores for two parallel, 20-item mathematics tests given to random samples from a national population of examinees.

Table 1 about here

It is evident that test Y, with a mean of 11.6 (±.1) is about one raw score point <u>easier</u> than test X, which has a mean of 10.8 (±.1). In this example, the single zero in the Y-frequencies would prevent the raw sample proportions from satisfying the positivity condition. Table 2 shows the fitted frequencies and Freeman-Tukey residuals (Bishop, Fienberg and Holland, 1975) for log-linear models of the form

$$\log r_j = \alpha + \sum_{i=1}^{L_X} \beta_i (x_j)^i$$

and (29)

$$\log s_k = \alpha' + \sum_{i=1}^{L_Y} \beta_i' (y_k)^i,$$

with $L_X = 2$ and $L_Y = 3$. The likelihood ratio chi-square statistic for the model for $\{r_j\}$ is 18.35 on 18 degrees of freedom while that for $\{s_k\}$ is 20.24 on 17 degrees of freedom and these values suggest that, overall, the fits of these two models are quite good. To get a more detailed look at these fits we examine the Freeman-Tukey residuals in Table 2. These residuals should behave roughly like independent standard normal deviates if the model fits adequately. Since these residuals all lie within ± 2.0 and show no pattern we conclude that the fitted probabilities (i.e. $\hat{r_j}$ and $\hat{s_k}$) from these models are improved estimates of the population score distributions in the sense of "consistency" and "stability" described earlier.

Table 2 about here

Table 1

Score Frequencies for Tests X and Y for Random Samples from the Same Population

Score	X-frequencies	Y-frequencies
0	1	0
1	3	4
2	8	11
3	25	16
4	30	18
5	64	34
6	67	63
7	95	89
8	116	87
9	124	129
10	156	124
11	147	154
12	120	125
13	129	131
14	110	109
15	86	98
16	66	89
17	51	66
18	29	54
19	15	37
20	11	17
Total	1453	1455
Mean	10.8	11.6
Sd	3.8	3.9

Table 2

Fitted Score Frequencies and Freeman-Tukey Residuals for Tests X and Y for Random Samples From the Same Populations

Score	Test X Fitted Frequencies*	FT Residuals	Test Y Fitted Frequencies**	FT Residuals
0	3.30	-1.4	1.71	-1.8
1	6.44	-1.4	3.77	0.2
2	11.77	-1.1	7.65	1.2
3	20.17	1.1	14.24	0.5
4	32.43	-0.4	24.44	-1.3
5	48.89	2.0	38.75	-0.7
6	69.10	-0.2	56.98	0.8
7	91.57	0.4	77.91	1.2
8	113.79	0.2	99.35	-1.3
9	132.58	-0.7	118.54	1.0
10	144.83	0.9	132.72	-0.8
11	148.36	-0.1	139.87	1.2
12	142.49	-1.9	139.15	-1.2
13	128.32	0.1	131.10	0.0
14	108.35	0.2	117.31	-0.8
15	85.79	0.1	100.00	-0.2
16	63.69	0.3	81.46	0.8
17	44.33	1.0	63.60	0.3
18	28.93	0.1	47.73	0.9
19	17.71	-0.6	34.54	0.5
20	10.16	0.3	24.18	-1.5

*2-moment fit

**3-moment fit

3.2 The Anchor-Test Equating Design

In Angoff's Design IV class of equating experiments, two independent random samples are drawn from two different populations, P and Q. Test X and an anchor-test, A, are given to the P-sample, while test Y and the anchor-test, A, is given to the Q-sample. Angoff's Design III is similar except that, in Design III, P and Q are the same population.

In the anchor-test designs, when P and Q differ, there is a choice of population on which to do the equating. In general the synthetic population, S, describes this choice of populations. Let w be a proportion, $0 \leq w \leq 1$, then S may be denoted $wP + (1-w)Q$ and viewed as composed of two strata, P and Q, that are given relative weight w and 1-w, respectively. This means that probabilities for S are defined as weighted averages of corresponding P and Q probabilities. For example, $\text{Prob}_S\{X = x_j\}$ is defined by:

$$r_j = \text{Prob}_S\{X = x_j\} = w\text{Prob}_P\{X = x_j\} + (1-w)\text{Prob}_Q\{X = x_j\}$$

and (30)

$$s_k = \text{Prob}_S\{Y = y_k\} = w\text{Prob}_P\{Y = y_k\} + (1-w)\text{Prob}_Q\{Y = y_k\}.$$

However, (30) shows the need to estimate probabilities for which there can be no data, i.e., $\text{Prob}_Q\{X = x_j\}$ and $\text{Prob}_P\{Y = y_k\}$. This estimation must be accomplished by making assumptions that, in general, can not be tested. One such assumption, originally suggested by Tucker and discussed in Braun and Holland (1982) is the what we call the Conditional Homogeneity Assumption defined below:

Conditional Homogeneity Assumption: The conditional distribution of X given A (and of Y given A) is the same (i.e. is homogeneous) in P and Q, i.e.,

$$\text{Prob}_P\{X = x_j | A = a_u\} = \text{Prob}_Q\{X = x_j | A = a_u\}$$

and (31)

$$\text{Prob}_P\{Y = y_k | A = a_u\} = \text{Prob}_Q\{Y = y_k | A = a_u\}.$$

Note that when P = Q the conditional homogeneity assumption is automatically satisfied.

We call the assumption "conditional homogeneity" because it asserts that the conditional distributions of X (and of Y) in P and Q are homogeneous, i.e. the same in the two populations.

The next theorem summarizes the use of this assumption in the estimation or calculation of $\text{Prob}_Q\{X = x_j\}$ and $\text{Prob}_P\{Y = y_k\}$.

Theorem 4: Under the Conditional Homogeneity Assumption

$$\text{Prob}_Q\{X = x_j\} = \sum_u \text{Prob}_P\{X = x_j | A = a_u\} \text{Prob}_Q\{A = a_u\}$$

and (32)

$$\text{Prob}_P\{Y = y_k\} = \sum_u \text{Prob}_Q\{Y = y_k | A = a_u\} \text{Prob}_P\{A = a_u\}.$$

The proof of this result is straight-forward and omitted.

In (32) we see that the right-hand sides of the equations involve only parameters (i.e. probabilities) that can, in principle, be estimated from the data collected in the design. When (32) is combined with (30), the probabilities $\{r_j\}$ and $\{s_k\}$ can all be estimated. The relevant equations on which this estimation is based are given below:

$$r_j = w\text{Prob}_P\{X = x_j\} + (1-w) \sum_u \text{Prob}_P\{X = x_j | A = a_u\} \text{Prob}_Q\{A = a_u\},$$

and (33)

$$s_k = (1-w)\text{Prob}_Q\{Y = y_k\} + w \sum_u \text{Prob}_Q\{Y = y_k | A = a_u\} \text{Prob}_P\{A = a_u\}.$$

The raw data that arises in anchor-test designs consists of two sets of bivariate frequencies, i.e., the (X,A)-frequencies from P,

$$n_{ju} = \text{number of examinees with } X = x_j \text{ and } A = a_u$$

and the (Y,A)-frequencies from Q,

$$m_{ku} = \text{number of examinees with } Y = y_k,\ A = a_u.$$

The two sample sizes are given by

$$n = \sum_{j,u} n_{ju} \quad \text{and} \quad m = \sum_{k,u} m_{ku}.$$

The raw sample frequencies could be used to estimate the various probabilities that go to make r_j and s_k given in (33). However, rarely will these raw sample frequencies yield satisfactory estimates of all the probabilities involved except when m and n are very large. Tables 3 and 4 give bivariate frequencies for (X,A) and (Y,A) where X and Y are the same as in section 3.1 and A is a 20 item anchor-test that is parallel to X and Y. Note that in this example, P = Q so that the conditional homogeneity assumption is automatically satisfied.

Tables 3 and 4 about here

Let $\{p_{ju}\}$ and $\{q_{ku}\}$ be the population joint distribution given by

$$p_{ju} = \text{Prob}_P\{X = x_j,\ A = a_u\}$$

$$q_{ku} = \text{Prob}_Q\{Y = y_k,\ A = a_u\}. \tag{34}$$

Tables 5 and 6 give the fitted distributions that are obtained by fitting log-linear models of the form

Table 3

Bivariate Score Distribution for Tests X and A

X Score	Anchor Score A																				
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	XMARGIN
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
2	0	0	1	0	2	1	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	8
3	0	0	1	5	6	3	8	1	1	0	0	0	0	0	0	0	0	0	0	0	0	25
4	0	0	2	7	4	6	4	3	1	3	0	0	0	0	0	0	0	0	0	0	0	30
5	0	0	3	3	5	12	14	8	9	6	2	1	0	0	0	0	0	0	0	0	0	64
6	0	0	1	4	10	9	12	9	8	10	4	0	0	0	0	0	0	0	0	0	0	67
7	0	0	1	3	5	7	16	16	11	17	10	5	3	0	1	0	0	0	0	0	0	95
8	0	0	1	1	3	8	16	14	12	24	20	11	3	3	0	0	0	0	0	0	0	116
9	0	0	0	1	3	4	8	19	20	17	17	13	11	9	2	0	0	0	0	0	0	124
10	0	0	0	0	1	2	6	14	20	19	28	24	17	11	9	3	2	0	0	0	0	156
11	0	0	0	0	1	3	3	6	13	17	21	23	27	14	13	2	2	1	1	0	0	147
12	0	0	0	0	0	1	0	5	11	14	16	26	18	11	10	3	3	1	1	0	0	120
13	0	0	0	0	0	0	1	4	8	8	20	21	19	16	13	9	6	3	1	0	0	129
14	0	0	0	0	0	0	0	1	4	3	3	17	18	26	11	21	4	1	1	0	0	110
15	0	0	0	0	0	0	1	0	1	3	4	10	12	15	15	10	10	3	1	1	0	86
16	0	0	0	0	0	0	0	0	0	1	1	1	11	12	8	13	10	7	1	1	0	66
17	0	0	0	0	0	0	0	0	0	0	2	1	5	4	8	9	11	5	3	3	0	51
18	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	4	4	11	4	1	0	29
19	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	2	2	3	3	1	0	15
20	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	3	3	2	0	11
																					N =	1453

Table 4

Bivariate Score Distribution of Tests Y and A

YSCORE	Anchor Score A																				
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	YMARGIN
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	4
2	0	0	1	4	4	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	11
3	0	0	2	2	5	4	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	16
4	0	3	0	3	1	5	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	18
5	0	0	0	4	6	5	8	3	4	3	0	0	0	0	0	1	0	0	0	0	0	34
6	0	0	0	3	5	9	8	20	9	5	0	3	0	1	0	0	0	0	0	0	0	63
7	0	0	0	2	7	6	15	13	17	12	10	3	2	2	0	0	0	0	0	0	0	89
8	0	0	0	1	3	10	11	16	15	10	7	8	3	3	0	0	0	0	0	0	0	97
9	0	0	0	0	5	5	9	18	23	21	20	14	6	5	2	1	0	0	0	0	0	129
10	0	0	0	0	1	2	8	15	9	26	23	15	14	5	2	3	0	1	0	0	0	124
11	0	0	0	0	2	5	10	9	22	19	32	20	9	15	6	2	2	1	0	0	0	154
12	0	0	0	1	0	0	1	6	9	18	20	19	13	20	11	4	3	0	0	0	0	125
13	0	0	0	0	0	0	5	2	4	13	14	22	19	26	11	9	5	1	0	0	0	131
14	0	0	0	0	0	0	1	2	4	9	12	14	20	11	13	12	8	2	1	0	0	109
15	0	0	0	0	0	0	0	0	3	1	6	10	20	16	15	12	7	7	0	1	0	98
16	0	0	0	0	0	0	0	0	0	0	3	13	13	5	16	16	8	11	2	2	0	89
17	0	0	0	0	0	0	0	0	0	1	2	0	4	10	11	11	13	9	5	0	0	66
18	0	0	0	0	0	0	0	0	0	0	0	1	1	5	6	7	20	7	4	3	0	54
19	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	7	9	14	4	0	0	37
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	7	3	4	1	17
																					N =	1455

$$\log (p_{ju}) = \alpha + \sum_{i=1}^{2} \beta_i (x_j)^i + \sum_{i=1}^{2} \gamma_i (a_u)^i + \delta\, x_j\, a_u$$

and (35)

$$\log (q_{ku}) = \alpha^* + \sum_{i=1}^{2} \beta_i^* (y_k)^i + \sum_{i=1}^{2} \gamma_i^* (a_u)^i + \delta^*\, y_k\, a_u \ .$$

Tables 5 and 6 about here

The likelihood ratio tests for adding extra terms to the models in (35) were not significant. Table 7 gives the estimates of $\hat{r}_j$ and $\hat{s}_k$ that follow from these smoothed distributions using (33), with w = .5.

This is an example of an external anchor test. In Holland, King and Thayer (1988) the internal anchor test is also discussed and shown to be easily transfomed to the external anchor-test case.

Table 7 about here

Table 5

Fitted Bivariate Score Distribution for Tests X and A

X SCORE	Anchor Score A																			
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0.34	0.49	0.59	0.58	0.47	0.32	0.18	0.08	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.45	0.74	1.02	1.15	1.08	0.84	0.54	0.29	0.13	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.50	0.95	1.49	1.94	2.10	1.87	1.38	0.84	0.43	0.18	0.06	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.47	1.02	1.85	2.77	3.44	3.53	2.99	2.10	1.22	0.59	0.23	0.08	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.37	0.93	1.95	3.35	4.78	5.64	5.50	4.44	2.97	1.64	0.75	0.28	0.09	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.25	0.72	1.73	3.43	5.62	7.62	8.55	7.94	6.10	3.88	2.04	0.89	0.32	0.10	0.02	0.00	0.00	0.00	0.00	0.00	0.00
6	0.14	0.47	1.30	2.97	5.60	8.73	11.26	12.03	10.63	7.77	4.70	2.35	0.97	0.33	0.09	0.02	0.00	0.00	0.00	0.00	0.00
7	0.07	0.28	0.83	2.18	4.72	8.46	12.56	15.42	15.66	13.16	9.16	5.27	2.51	0.99	0.32	0.09	0.02	0.00	0.00	0.00	0.00
8	0.03	0.12	0.45	1.35	3.37	6.95	11.85	16.73	19.54	18.89	15.11	10.00	5.48	2.48	0.93	0.29	0.07	0.02	0.00	0.00	0.00
9	0.01	0.05	0.21	0.71	2.04	4.83	9.47	15.38	20.65	22.95	21.10	16.06	10.11	5.27	2.27	0.81	0.24	0.06	0.01	0.00	0.00
10	0.00	0.02	0.08	0.32	1.04	2.84	6.41	11.96	18.47	23.61	24.96	21.84	15.81	9.47	4.70	1.93	0.65	0.18	0.04	0.01	0.00
11	0.00	0.00	0.03	0.12	0.45	1.42	3.67	7.88	13.99	20.56	25.00	25.14	20.93	14.42	8.22	3.88	1.51	0.49	0.13	0.03	0.01
12	0.00	0.00	0.01	0.04	0.17	0.60	1.78	4.40	8.97	15.18	21.19	24.51	23.46	18.58	12.18	6.60	2.96	1.10	0.34	0.09	0.02
13	0.00	0.00	0.00	0.01	0.05	0.21	0.73	2.08	4.87	9.47	15.21	20.23	22.27	20.28	15.28	9.53	4.92	2.10	0.74	0.22	0.05
14	0.00	0.00	0.00	0.00	0.01	0.06	0.25	0.83	2.24	5.00	9.25	14.14	17.89	18.73	16.23	11.64	6.90	3.39	1.38	0.46	0.13
15	0.00	0.00	0.00	0.00	0.00	0.02	0.07	0.28	0.87	2.24	4.76	8.37	12.17	14.66	14.60	12.04	8.21	4.63	2.16	0.84	0.27
16	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.08	0.29	0.85	2.07	4.19	7.01	9.71	11.12	10.54	8.27	5.36	2.88	1.28	0.47
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.08	0.27	0.77	1.78	3.42	5.44	7.17	7.81	7.05	5.26	3.25	1.66	0.70
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.07	0.24	0.64	1.41	2.59	3.91	4.91	5.09	4.38	3.10	1.82	0.88
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.06	0.19	0.49	1.04	1.81	2.61	3.11	3.07	2.50	1.69	0.95
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.05	0.15	0.35	0.71	1.17	1.61	1.82	1.71	1.33	0.85

BEST COPY AVAILABLE

Table 6

Fitted Bivariate Score Distribution for Test Y and A

Y SCORE	Anchor Score A																				
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0.33	0.44	0.49	0.45	0.34	0.22	0.11	0.05	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.45	0.68	0.87	0.92	0.81	0.59	0.35	0.18	0.07	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.51	0.90	1.32	1.61	1.82	1.35	0.94	0.54	0.26	0.10	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.50	1.02	1.71	2.39	2.76	2.65	2.11	1.40	0.76	0.35	0.13	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.42	0.97	1.88	3.02	4.02	4.44	4.06	3.09	1.94	1.01	0.44	0.16	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.30	0.79	1.77	3.28	4.99	6.33	6.66	5.81	4.21	2.52	1.26	0.52	0.18	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00
6	0.18	0.55	1.41	2.99	5.27	7.68	9.29	9.32	7.76	5.35	3.06	1.45	0.57	0.19	0.05	0.01	0.00	0.00	0.00	0.00	0.00
7	0.09	0.33	0.98	2.34	4.74	7.95	11.05	12.75	12.19	9.67	6.36	3.47	1.57	0.59	0.18	0.05	0.01	0.00	0.00	0.00	0.00
8	0.04	0.16	0.56	1.56	3.63	7.01	11.20	14.85	16.32	14.88	11.25	7.05	3.67	1.58	0.57	0.17	0.04	0.01	0.00	0.00	0.00
9	0.02	0.07	0.28	0.89	2.37	5.28	9.67	14.74	18.62	19.52	16.96	12.22	7.31	3.62	1.49	0.51	0.14	0.03	0.01	0.00	0.00
10	0.00	0.03	0.12	0.43	1.32	3.37	7.11	12.48	18.10	21.81	21.79	18.05	12.40	7.07	3.34	1.31	0.43	0.11	0.03	0.00	0.00
11	0.00	0.01	0.04	0.18	0.63	1.84	4.46	8.98	14.99	20.76	23.85	22.71	17.94	11.75	6.38	2.87	1.07	0.33	0.09	0.02	0.00
12	0.00	0.00	0.01	0.06	0.25	0.85	2.38	5.51	10.58	16.84	22.24	24.35	22.10	16.64	10.39	5.38	2.31	0.82	0.24	0.06	0.01
13	0.00	0.00	0.00	0.02	0.09	0.34	1.08	2.88	6.36	11.64	17.67	22.24	23.21	20.09	14.42	8.58	4.24	1.73	0.59	0.17	0.04
14	0.00	0.00	0.00	0.00	0.03	0.11	0.42	1.28	3.26	6.88	11.96	17.30	20.76	20.66	17.04	11.66	6.62	3.11	1.22	0.39	0.11
15	0.00	0.00	0.00	0.00	0.01	0.03	0.14	0.49	1.42	3.44	6.90	11.47	15.83	18.10	17.17	13.50	8.81	4.78	2.14	0.79	0.26
16	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.16	0.53	1.47	3.39	6.48	10.28	13.51	14.73	13.32	9.99	6.21	3.20	1.37	0.49
17	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.17	0.54	1.42	3.12	5.69	8.60	10.77	11.20	9.65	6.90	4.09	2.01	0.82
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.05	0.17	0.51	1.28	2.68	4.66	6.71	8.02	7.95	6.53	4.45	2.51	1.18
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.15	0.45	1.08	2.15	3.58	4.89	6.57	5.28	4.12	2.88	1.44
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.13	0.37	0.85	1.61	2.54	3.33	3.28	3.28	2.43	1.51

Table 7

Estimated Values for $\{r_j\}$ and $\{s_k\}$ Computed from Equation (33) and the Fitted Distributions in Tables 5 and 6

X SCORE	PROBABILITIES	Y SCORE	PROBABILITIES
0	0.002	0	0.002
1	0.004	1	0.004
2	0.008	2	0.006
3	0.014	3	0.011
4	0.022	4	0.018
5	0.033	5	0.028
6	0.046	6	0.039
7	0.061	7	0.053
8	0.076	8	0.067
9	0.089	9	0.080
10	0.098	10	0.091
11	0.101	11	0.097
12	0.098	12	0.098
13	0.089	13	0.093
14	0.077	14	0.083
15	0.062	15	0.070
16	0.047	16	0.056
17	0.033	17	0.042
18	0.022	18	0.030
19	0.013	19	0.020
20	0.008	20	0.012

4. THE CONTINUIZATION STEP

There are a variety of ways to select the continuization constants h_X and h_Y. Perhaps the easiest is to always use specific fixed values such as $h_X = h_Y = \infty$, which corresponds to linear equating, or $h_X = h_Y = .3$, which we have shown to correspond roughly to traditional equipercentile equating. Rather than always using fixed choices of h_X and h_Y, we suggest a flexible approach toward the choice of continuization constants, remembering that various goals may need to be achieved in selecting a satisfactory equating function.

Our approach is to choose h_X so that $F_{h_X}(x)$ is close to $F(x)$ in some sense. Some care needs to be exercised in selecting a notion of closeness. For example, if the sup norm, i.e.,

$$\sup_x |F_{h_X}(x) - F(x)| \tag{36}$$

is used to measure how close $F_{h_X}(x)$ is to $F(x)$, then this is minimized for $h_X = 0$ and the result is useless.

The density of $F_{h_X}(x)$, i.e. $F'_{h_X}(x)$, can be used to clarify what we want in a "good" continuous approximation to $F(x)$. Consider Figure 1. It is the density that arises when $h_X = .3$ in the example of section 3.1. It exhibits a "stegosaurian" character that would appear, on its face, to be undesirable. When $h_X = 1.0$, the result is Figure 2. Evidently, h_X has a big influence on the shape of the continuous approximation for $F(x)$.

When the x_j are consecutive integers, we can use the density, $F'_{h_X}(x)$, to create a histogram that we can then compare to the $\{\hat{r}_j\}$. This is done in the following way. Imagine a histogram centered on the $\{x_j\}$ with heights $\{F'_{h_X}(x_j)\}$ and unit width. If h_X is chosen appropriately this histogram will be close to the unit width histogram on the x_j with heights $\{\hat{r}_j\}$. To choose h_X we can

---

Figures 1 and 2 about here

Figure 1

Graph of the Density $F'_{h_X}(x)$ for $h_X = .3$

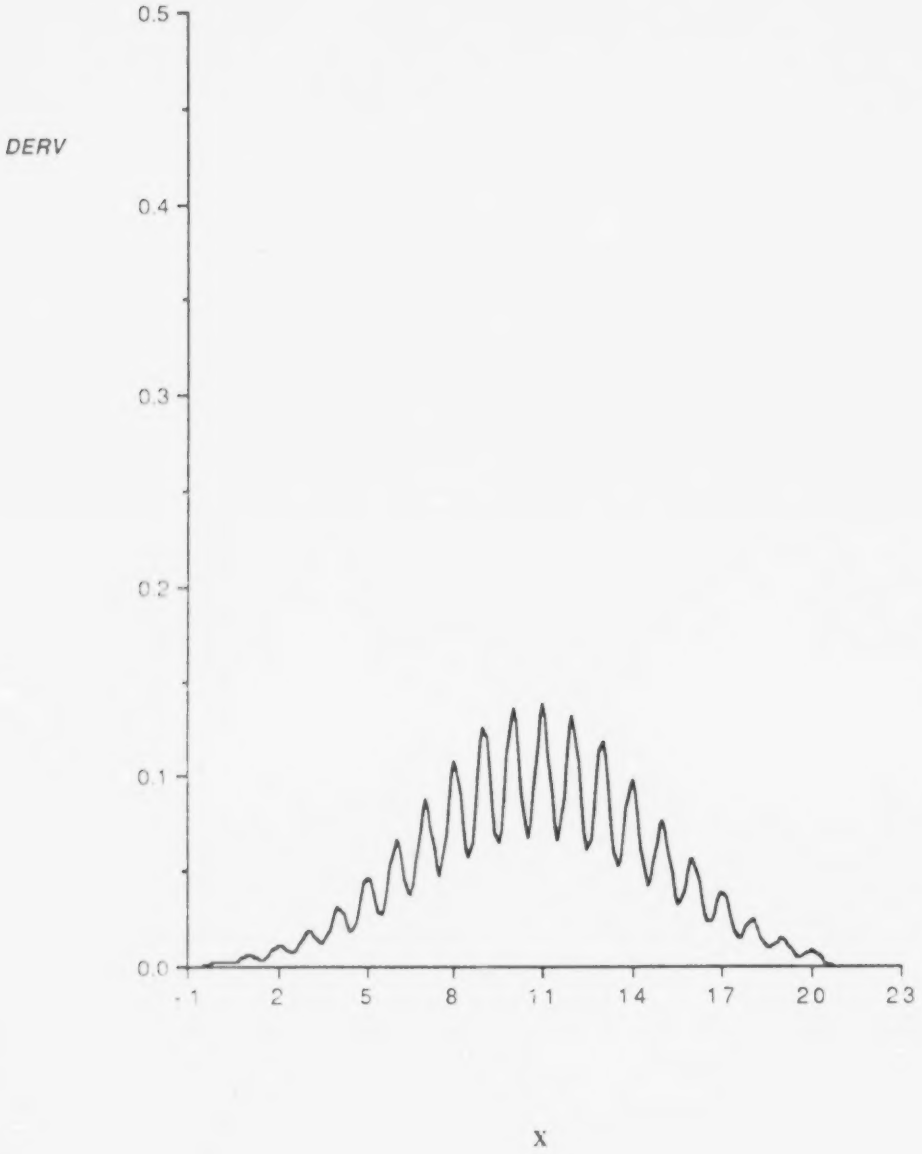
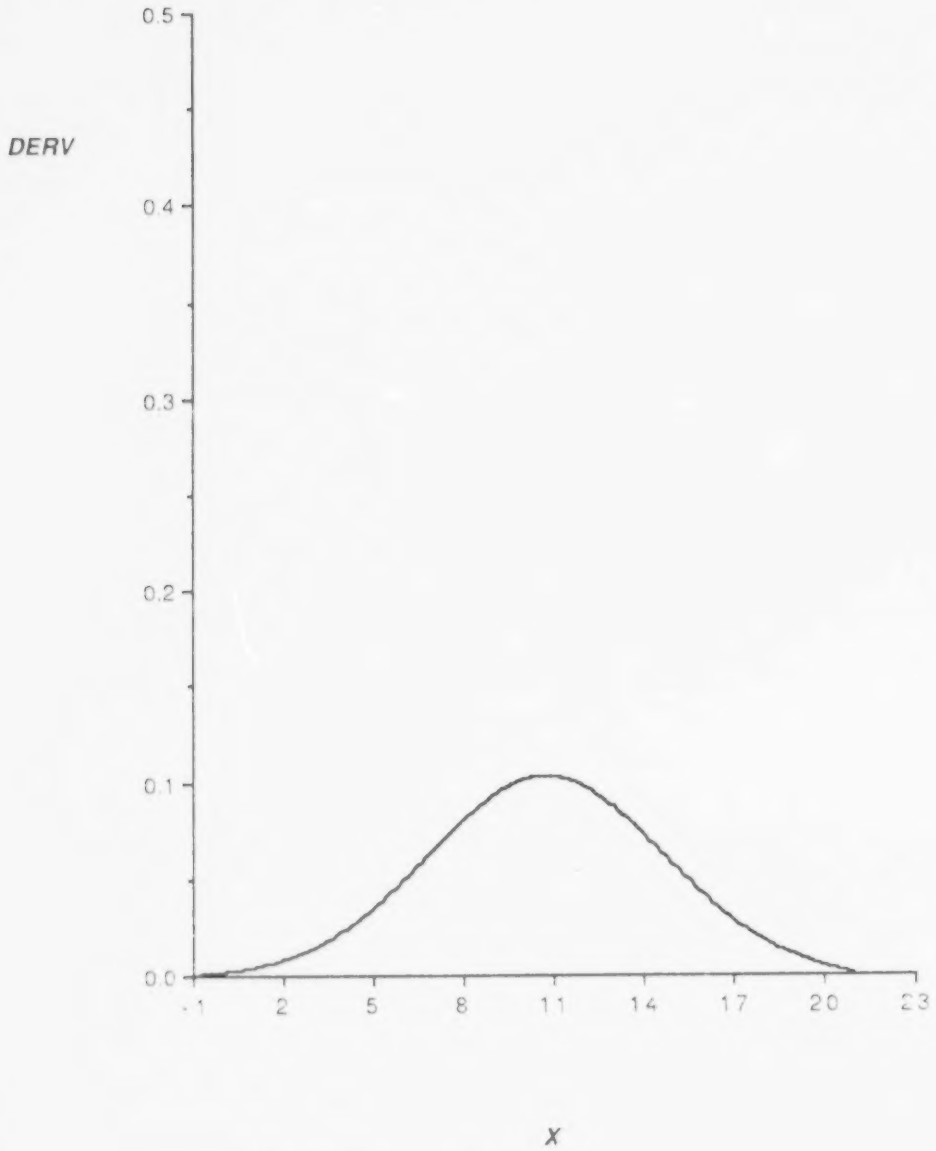


Figure 2

Graph of the Density $F'_{h_X}(x)$ for $h_X = 1.0$



minimize the "squared difference" criterion.

$$\sum_j (\hat{r}_j - F_{h_X}(x_j))^2 \ . \qquad (37)$$

The minimizing values of h_X and h_Y for the example of section 3.1 are .62 and .57 respectively.

In the case of anchor test equating, i.e. section 3.2, the same considerations arise but are applied to $\{r_j\}$ and $\{s_k\}$ from (33). Using the estimates of r_j and s_k in Table 7, the optimal values of h_X and h_Y that minimize the squared difference criterion are .62 and .59, respectively.

The continuization step can be used to remove the need for a final "postsmoothing" of the equating function (Fairbank (1985), Kolen (1984), Kolen and Jarjoura (1987)). The reason postsmoothing arises is that if the continuous approximations to F and G are not smooth enough, the equating functions computed via (11) will exhibit unreasonable oscillations about an otherwise smooth trend. Postsmoothing eliminates these oscillations. One situation that can produce these oscillations arises when tests are formula-scored. In formula-scored tests with few omitted responses the raw-score distribution will often produce "gaps" at specific scores. Figure 3 illustrates this phenomenon. When smoothing frequencies that exhibit gaps one has the choice of whether or not the smoothed frequencies ought to have "gaps" in them. Figure 4 shows a fitted distribution to the data in Figure 3 that has gaps. It was achieved by fitting moments to the "gap" scores as well as to all the scores using the techniques discussed in Holland and Thayer (1987). If a distribution that had no gaps had been fit to these data, the fit would have been poor according to the usual

---

Figures 3 and 4 about here

Figure 3

A Raw-score Distribution for a Formula-scored Test
That Exhibits "Gaps" at Regular Intervals on the Score Scale

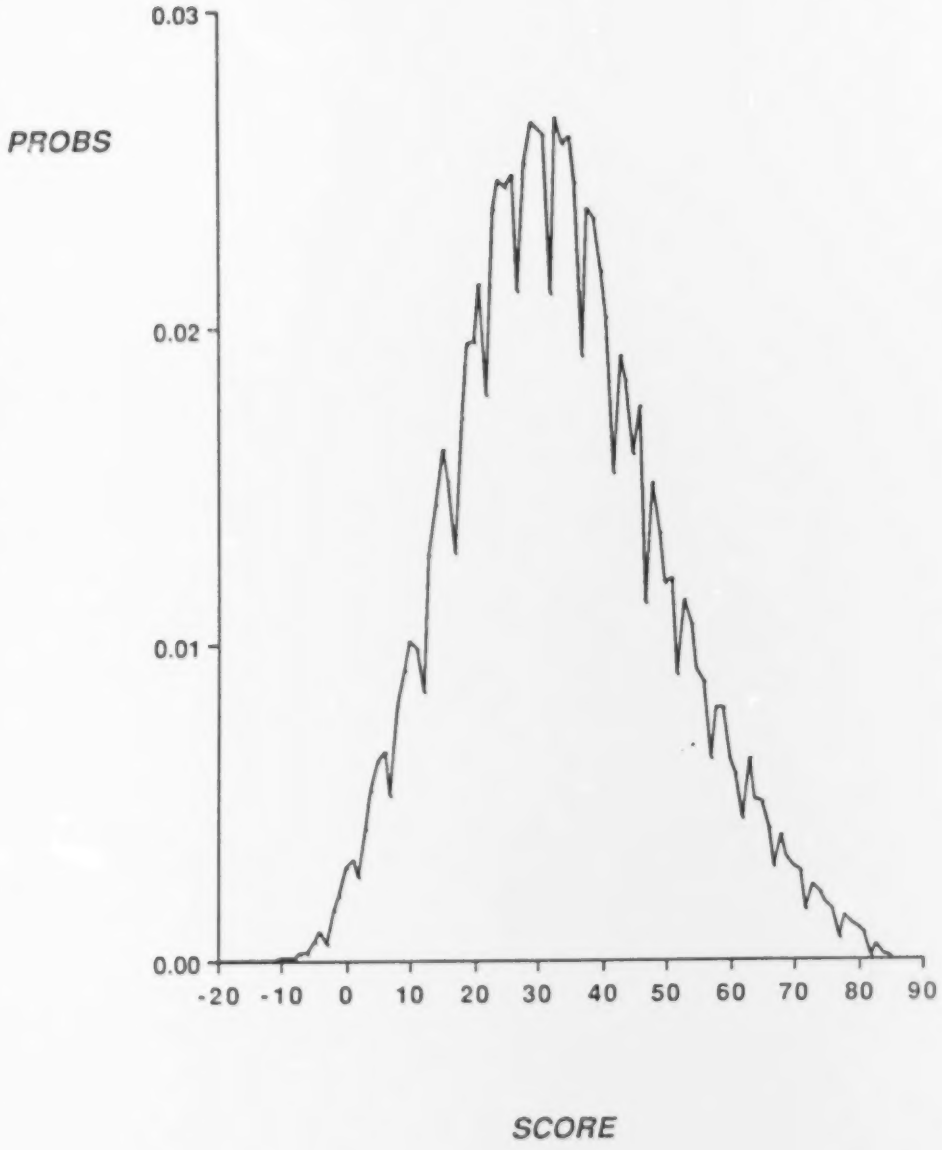
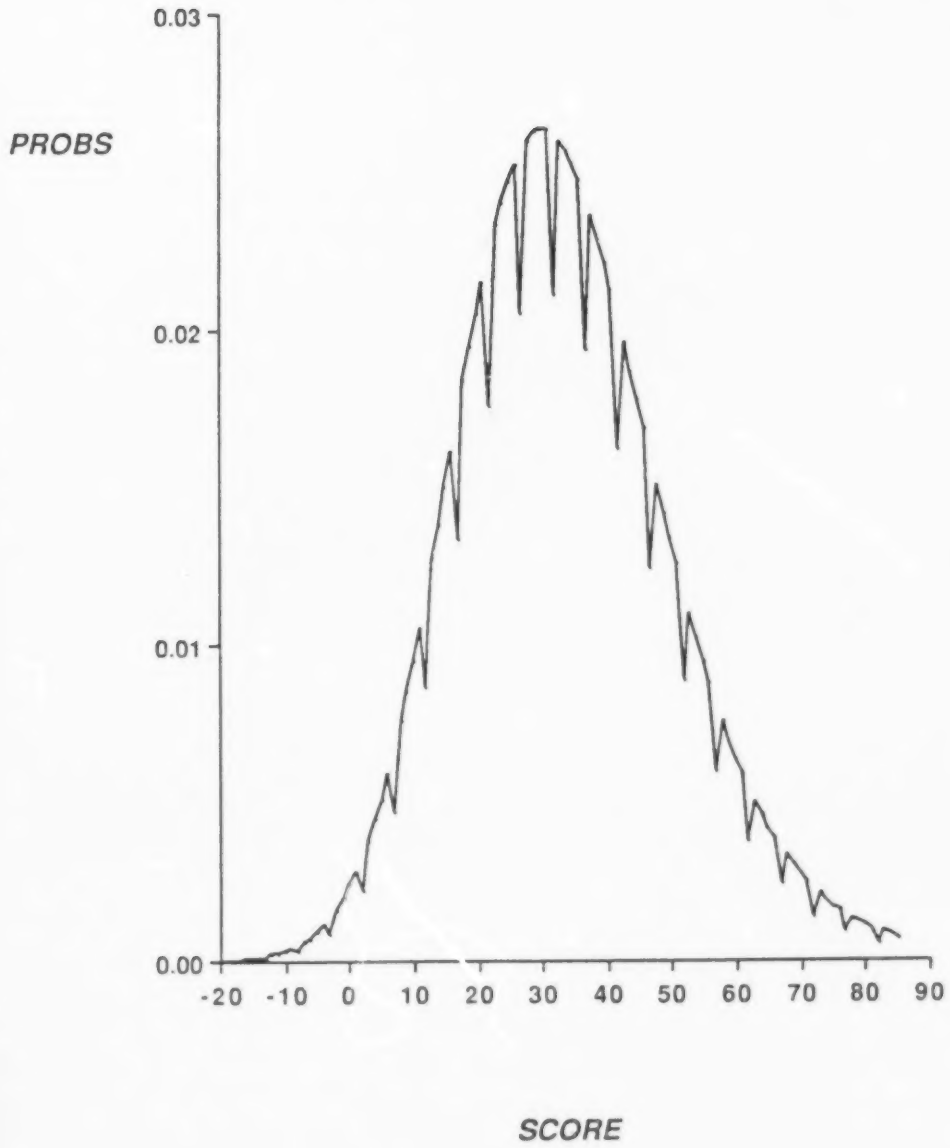


Figure 4

A Model With "Gaps" Fitted to the Data in Figure 3



goodness-of-fit statistics and it would have been unclear how to choose a satisfactory model. When data with gaps are encountered in test equating we recommend that the gaps be accounted for in the estimation step, i.e. by fitting a model like the one in Figure 4. The reason is that standard goodness-of-fit tests then provide a rational basis for choosing a model, and the resulting estimated standard errors for the fitted model (used to compute the standard error of equating) can be expected to be approximately correct. In the continuization step, the gaps can then be removed by taking h_X large enough. Figures 5 and 6 show the approximating densities for the fitted model in Figure 4 for $h_X = 1$ and 3, respectively. When $h_X = 1$ there are still some remnants of the gaps left but by $h_X = 3$ they are gone and the undesirable oscillations have been smoothed out. Figure 7 shows the fitted probabilities from Figure 4 and the continuous density for $h_X = 3$ from Figure 6. The density shows the general shape of the fitted probabilities but the gaps have been filled in.

We recommend that gaps be preserved in the estimation step and then removed in the continuization step in order to insure the accuracy of the standard error of equating that is discussed extensively in the companion paper, Holland, King and Thayer (1988).

Figures 5, 6 and 7 about here

Figure 5

Graph of the Density $F'_{h_X}(x)$ for $h_X = 1.0$ for $\{r_j\}$ in Figure 4

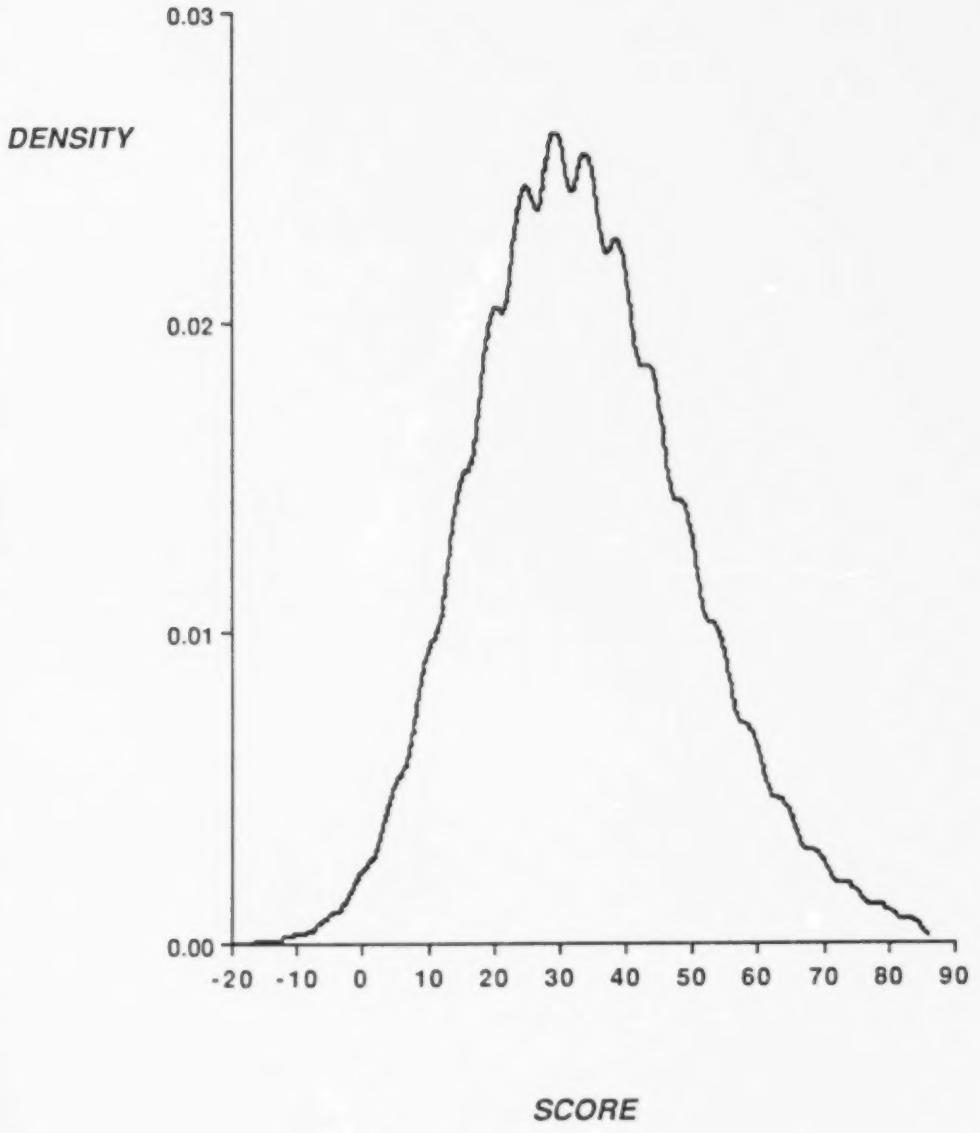


Figure 6

Graph of the Density $F'_{h_X}(x)$ for $h_X = 3.0$ for $\{r_j\}$ in Figure 4

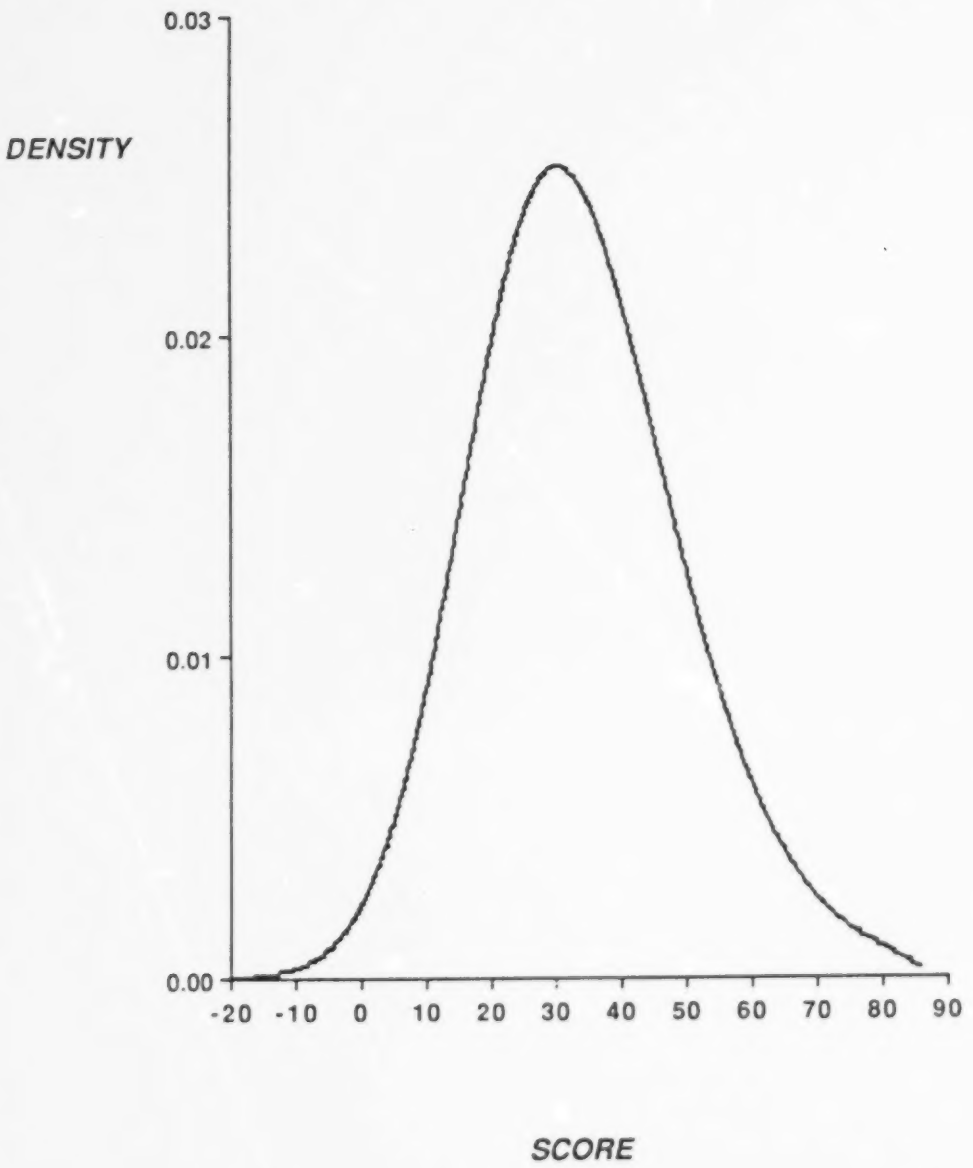
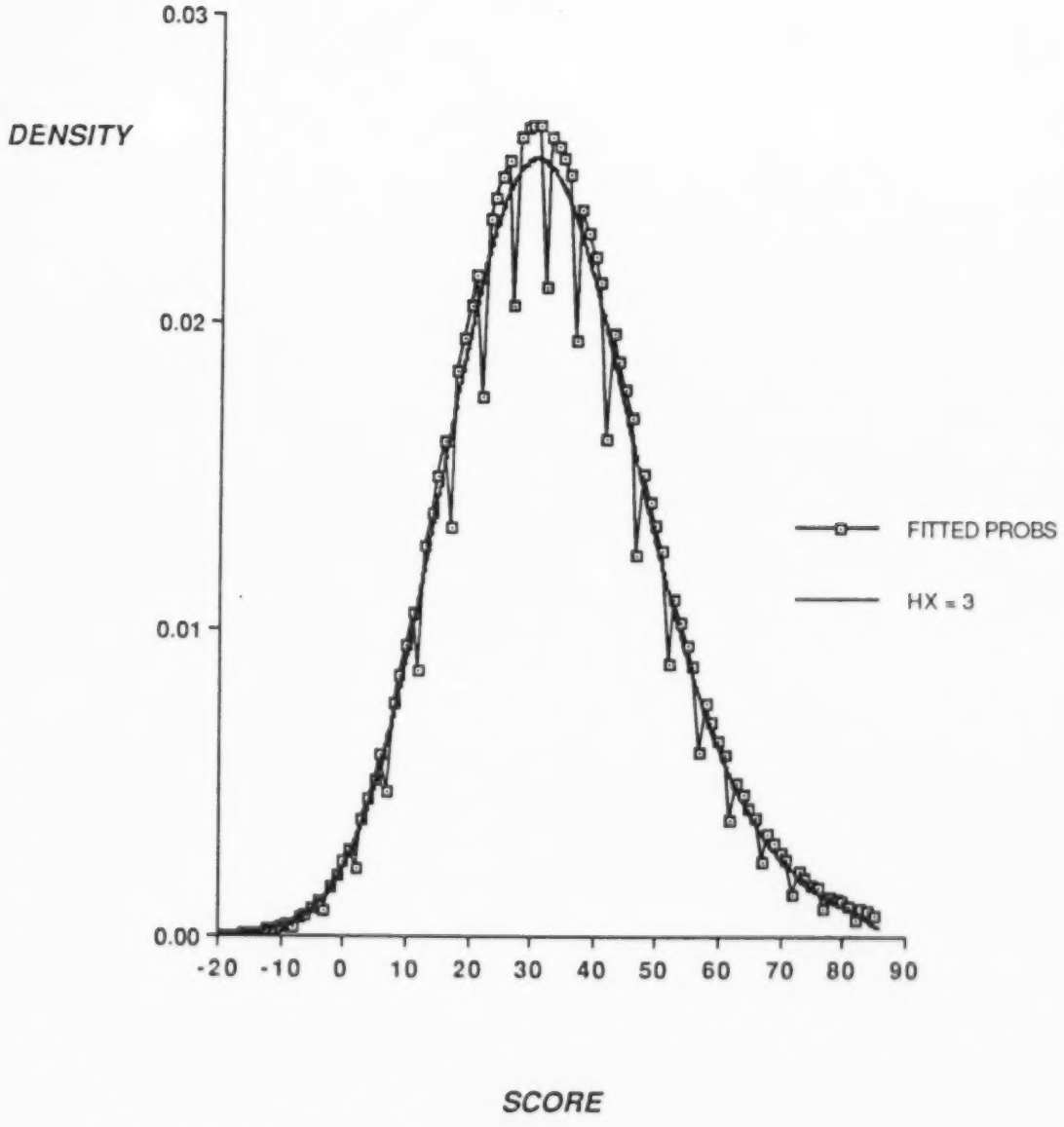


Figure 7

Graph of the Density for $h_X = 3.0$ and the Fitted Probabilities

Showing How the Gaps Have Been Filled In



5. THE EQUATING STEP

Once continuous approximations to $\hat{F}(x)$ and $\hat{G}(y)$ are in hand, it is a relatively straightforward process to compute the equating functions via (23) and (24). The only computational issue is the accuracy with which the inverse functions $F_{h_X}^{-1}(p)$ and $G_{h_Y}^{-1}(p)$ need to be approximated. We have not investigated this carefully but have found that for the cases we have considered a grid of width .05 has proved sufficient.

In the examples of sections 3.1 and 3.2 the equating functions are very nearly linear. Figure 8 shows the difference between the graphs of the linear equating function ($h_x = h_y = \infty$) and the approximate equipercentile equating function ($h_x = h_y = .3$) for equating Y to X for the example in section 3.1. While there are some differences between these equating functions they are quite small in this example. Figure 9 shows three equating functions for simulated data in which there is a great deal of curvilinearity when $h_X = h_Y = .3$. The equating functions for $h_X = h_Y = 5$ and $h_X = h_Y = 10$ are also shown to illustrate that as the h's increase the equating functions become more linear.

Once h_X and h_Y are selected, $F_{h_X}(x)$ and $G_{h_Y}(y)$ are determined as functions of the estimated score probabilities $\{\hat{r}_j\}$ and $\{\hat{s}_k\}$. The computation of the standard error of equating (SEE) can then proceed by a straight forward, but tedious, application of the δ-method of computing asymptotic variances of functions of random quantities -- in this case the random quantities are $\{\hat{r}_j\}$ and $\{\hat{s}_k\}$. This is the approach described in detail in our companion paper, Holland, King and Thayer (1988).

Figures 8 and 9 about here

Figure 8

The Difference Between the Linear and the Approximate Equipercentile Equating Functions, for the Example of Section 3.1

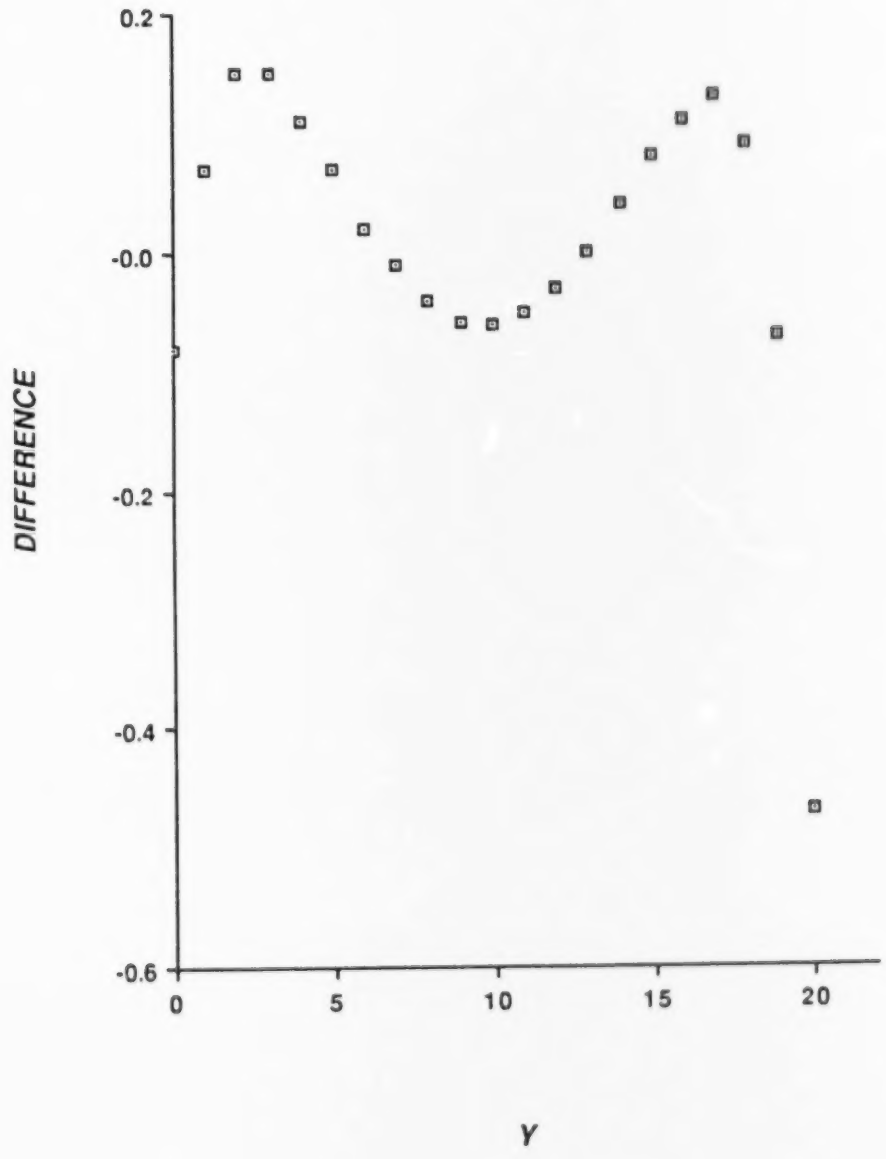
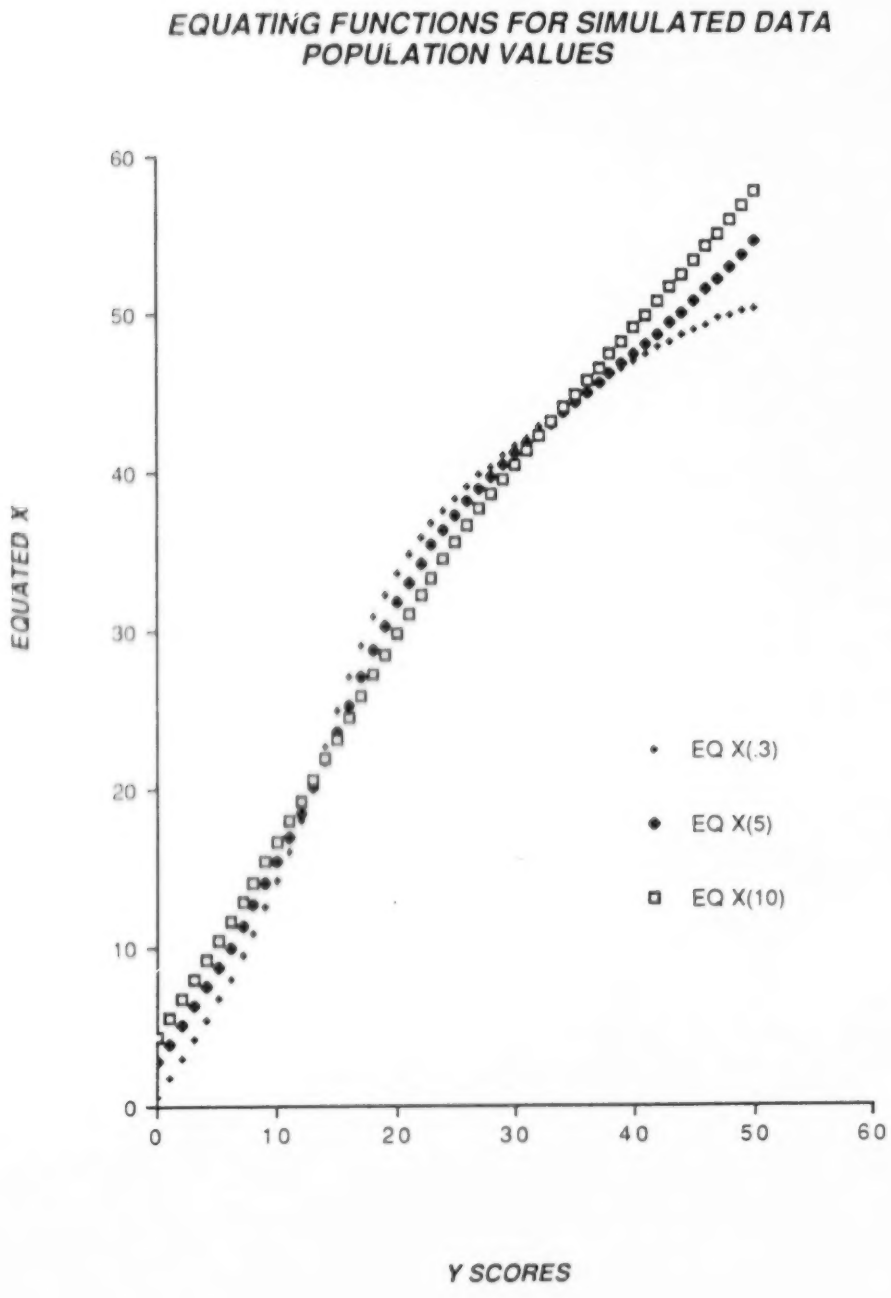


Figure 9

Three Equating Functions for Simulated Data



6. DISCUSSION

We believe that the kernel method of equating, when coupled with estimated score distributions using log-linear models, has a number of advantages over other observed-score equating methods.

First of all, the three phases, estimation, continuization and equating, form a unified approach to many problems that arise in equating. Most of the difficulties in equating arise in the estimation and continuization phases and these are quire different and ought to be treated separately. The problem of devising equating diagnostics is fairly easy once this separation is made. Some diagnostics will concern the estimation phase (i.e. the adequacy of model fit) while others concern the choice of continuization constant (e.g. the treatment of the "gaps" in formula score distributions).

Because log-linear models are very flexible they provide useful models for both large and small samples. Hence their use with the kernel method eliminates many of the problems that arise in equating with small samples of examinees. At the same time, large samples can also be fit adequately using these models.

The kernel method essentially contains linear and traditional equipercentile methods as special cases and can therefore exploit the best features of both methods. Furthermore, because it can handle both random groups and common item designs, the use of log-linear models in the latter case provides a substantially improved version of the method called "frequency estimation" (as called for in Braun and Holland, 1982).

The kernel method does not force the high and low score on the two tests to match as traditional equipercentile (and IRT true-score) methods do. It also does not restrict the equating function to be defined for only those raw score values that occur on the test. This can be very important for the chains of equatings that build up as a long sequence of new test forms is built up. In addition, because F_{h_X} and G_{h_X} are given by analytic formulas it is unecessary to specify the equating function by a table as most equipercentile methods do. Instead, if h_Y, h_X and the estimated probabillities $\{\hat{r}_j\}$ and $\{\hat{s}_k\}$ are kept, F_{h_X}, G_{h_Y} and the equating functions can be computed anew and chained together whenever they are needed. Although this is more complicated than carrying equating chains through by linear equating, it is still more satisfactory than the ad hoc tables of traditional equipercentile equating.

Finally, computationally efficient methods of estimating the standard error of equating are available and, for the first time, honest SEEs can be provided for a wide variety of equating designs. These SEEs reflect both the shape of the equating function, the design of the equating experiment, and the method used to pre-smooth the data in the estimation phase of the equating process.

In view of these advantages we see the kernel method of equating as a complete equating package that can provide measurement statisticians with a powerful set of tools for solving practical everyday problems in equating.

Future research in this area might explore a range of topics such as these.

1) Are there methods for choosing h_X and h_Y that are better than the minimization of the squared difference criterion, (37)?
2) What is the effect of data dependent choices of h_X and h_Y on the SEE?
3) Are the SEEs found by the δ-method good enough or are higher-order methods needed?
4) What is the relation between the kernel method and IRT or linear true-score equating methods?
5) What role can the kernel method play in the assessment of the invariance of equating functions across different populations of examinees?

REFERENCES

Angoff, W.H. (1984) Scales, norms and equivalent scores. Princeton, NJ: Educational Testing Service. (Reprinted from Thorndike, R.L. (Ed.) Educational Measurement, 1971).

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975) Discrete Multivariate Analysis: Theory and Practice. Cambridge, MA: MIT Press.

Braun, H.I. and Holland, P.W. (1982) Observed score test equating: a mathematical analysis of some ETS equating procedures. In Holland, P.W. and Rubin, D.B. (Eds.) Test Equating. New York: Academic Press.

Fairbank, B.A. (1985) Equipercentile test equating: the effects of presmoothing and postsmoothing on the magnitude of sample-dependent errors. (AFHRL-TR-84-64).

Holland, P.W. and Thayer, D.T. (1987) Notes on the use of log-linear models for fitting discrete probability distributions. Princeton, NJ: ETS Technical Report TR-87-79.

Holland, P.W., King, B.F. and Thayer, D.T. (1988) The standard error of equating for the kernel method of equating score distributions. Princeton, NJ: ETS Technical Report (in preparation).

Kendall, M. G. and Stuart, A. (1958) The Advanced Theory of Statistics, Vol. 1. London: Charles Griffin.

Kolen, M.J. (1984) Effectiveness of analytic smoothing in equipercentile equating. Journal of Educational Statistics, 9, 25-44.

Kolen, M.J. and Jarjoura, D. (1987) Analytic smoothing for equipercentile equating under the common item non-equivalent populations design. Psychometrika, 52, 43-60.

Lord, F.M. (1950) Notes on comparable scales for test scores. Princeton, NJ: ETS RB-50-48.

Rosenbaum, P.R. and Thayer, D.T. (1987) Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. British Journal of Mathematical and Statistical Psychology, 40, 43-49.

Tapia, R. A. and Thompson, J. R. (1978) Nonparametric Probability Density Estimation. Baltimore, MD: Johns Hopkins University Press.

APPENDIX: Proof of Theorem 2.

Let the moment generating function (mgf) of X be $M_X(t)$. It is well-known that the Taylor expansion of $\log[M_X(t)]$ is given by

$$\log[M_X(t)] = M_X t + \sigma_X^2 t^2/2 + \sum_{j\geq 3} k_{jX}(t)^j/j! \; . \tag{38}$$

But the mgf of $X(h_X)$ is given by

$$E[\exp\{tX(h_X)\}] =$$

$$E[\exp\{t(a_X(X+h_XV) + (1-a_X)\mu_X)\}]$$

$$= \exp\{t(1-a_X)\mu_X\}E[\exp\{ta_XX + ta_Xh_XV\}]$$

But since X and V are independent

$$E[\exp\{ta_XX + ta_Xh_XV\}]$$

$$= E[\exp\{ta_XX\}]E[\exp\{ta_Xh_XV\}]$$

$$= M_X(ta_X)\, M_V(ta_Xh_X)$$

where M_X and M_V are the mgfs of X and V respectively. But, it is well-known that

$$M_V(t) = \exp\{\tfrac{1}{2}\, t^2\},$$

so that the mgf of $X(h_X)$ can be expressed as

$$E[\exp\{tX(h_X)\}]$$

$$= \exp\{t(1-a_X)\mu_X\}\, M_X(ta_X)\, \exp\{\tfrac{1}{2}t^2a_X^2h_X^2\}.$$

Now take logs to get the cumulants, i.e.,

$$\log E[\exp\{tX(h_X)\}]$$

$$= t(1-a_X)\mu_X + \tfrac{1}{2}\, t^2a_X^2h_X^2 + \log[M_X(ta_X)]. \tag{39}$$

Now combine (38) and (39) to get

$$\log E[\exp\{tX(h_X)\}] =$$

$$((1-a_X)\mu_X + a_X\mu_X)t + (a_X^2h_X^2 + \sigma_X^2a_X^2)t^2/2$$

$$+ \sum_j k_{jX}(a_X)^j(t)^j/j!.$$

But $(1-a_X)\mu_X + a_X\mu_X = \mu_X$ and $a_X^2h_X^2 + \sigma_X^2a_X^2 = a_X^2(h_X^2 + \sigma_X^2) = \sigma_X^2$, so we obtain

$$\log E[\exp\{tX(h_X)\}] =$$

$$\mu_X t + \sigma_X^2 t^2/2 + \sum_{j\geq 3} (a_X)^j k_{jX}(t)^j/j!.$$

But the coefficients of a Taylor expansion are unique so the cumulants of $X(h_X)$ are $(a_X)^j k_{jX}$, QED.